Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models

Year: 2021 | Citations: 218 | Authors: Hannah Rose Kirk, Yennie Jun, Haider Iqbal, Elias Benussi, Filippo Volpin

Abstract

The capabilities of natural language models trained on large-scale data have increased immensely over the past few years. Open source libraries such as HuggingFace have made these models easily available and accessible. While prior research has identified biases in large language models, this paper considers biases contained in the most popular versions of these models when applied `out-of-the-box' for downstream tasks. We focus on generative language models as they are well-suited for extracting biases inherited from training data. Specifically, we conduct an in-depth analysis of GPT-2, which is the most downloaded text generation model on HuggingFace, with over half a million downloads per month. We assess biases related to occupational associations for different protected categories by intersecting gender with religion, sexuality, ethnicity, political affiliation, and continental name origin. Using a template-based data collection pipeline, we collect 396K sentence completions made by GPT-2 and find: (i) The machine-predicted jobs are less diverse and more stereotypical for women than for men, especially for intersections; (ii) Intersectional interactions are highly relevant for occupational associations, which we quantify by fitting 262 logistic models; (iii) For most occupations, GPT-2 reflects the skewed gender and ethnicity distribution found in US Labor Bureau data, and even pulls the societally-skewed distribution towards gender parity in cases where its predictions deviate from real labor market observations. This raises the normative question of what language models should learn - whether they should reflect or correct for existing inequalities.